PROCESS FOR MAINTAINING ONGOING REGISTRATION FOR PAGES ON A GIVEN SEARCH ENGINE

Matter enclosed in heavy brackets [ ] appears in the original patent but forms no part of this reissue specification; matter printed in italics indicates the additions made by reissue.

FIELD OF THE INVENTION

The present invention relates to the process of developing and maintaining the content of Internet search engine databases.

BACKGROUND OF THE INVENTION

An internet (including, but not limited to, the Internet, intranets, extranets and similar networks), is a network of computers, with each computer being identified by a unique address. The addresses are logically subdivided into domains or domain names (e.g. ibm.com, pbs.org, and oranda.net) which allow a user to reference the various addresses. A web, (including, but not limited to, the World Wide Web (WWW)) is a group of these computers accessible to each other via common communication protocols, or languages, including but not limited to Hypertext Transfer Protocol (HTTP). Resources on the computers in each domain are identified with unique addresses called Uniform Resource Locator (URL) addresses (e.g.http://www.ibm.com/products/laptops.htm). A web site is any destination on a web. It can be an entire individual domain, multiple domains, or even a single URL.

Resources can be of many types. Resources with a ".htm" or."html" URL suffix are text files, or pages, formatted in a specific manner called Hypertext Markup Language (HTML). HTML is a collection of tags used to mark blocks of text and assign meaning to them. A specialized computer application called a browser can decode the HTML files and display the information contained within. A hyperlink is a navigable reference in any resource to another resource on the internet.

An internet Search Engine is a web application consisting of

1. Programs which visit and index the web pages on the internet.
2. A database of pages that have been indexed
3. Mechanisms for a user to search the database of pages.

Agents are programs that can travel over the internet and access remote resources. The internet search engine uses agent programs called Spiders, Robots, or Worms, among other names, to inspect the text of resources on web sites. Navigable references to other web resources contained in a resource are called hyperlinks. The agents can follow these hyperlinks to other resources. The process of following hyperlinks to other resources, which are then indexed, and following the hyperlinks contained within the new resource, is called spidering.

The main purpose of an internet search engine is to provide users the ability to query the database of internet content to find content that is relevant to them. A user can visit the search engine web site with a browser and enter a query into a form (or page), including but not limited to an HTML form, provided for the task. The query may be in several different forms, but most common are words, phrases, or questions. The query data is sent to the search engine through a standard interface, including but not limited to the Common Gateway Interface (CGI). The CGI is a means of passing data between a client, a computer requesting data or processing and a program or script on a server, a computer providing data or processing. The combination of form and script is hereinafter referred to as a script application. The search engine will inspect its database for the URLs of resources most likely to relate to the submitted query. The list of URL results is returned to the user, with the format of the returned list varying from engine to engine. Usually it will consist of ten or more hyperlinks per search engine page, where each hyperlink is described and ranked for relevance by the search engine by means of various information such as the title, summary, language, and age of the resource. The returned hyperlinks are typically sorted by relevance, with the highest rated resources near the top of the list.

The World Wide Web consists of thousands of domains and millions of pages of information. The indexing and cataloging of content on an Internet search engine takes large amounts of processing power and time to perform. With millions of resources on the web, and some of the content on those resources changing rapidly (by the day, or even minute), a single search engine cannot possibly maintain a perfect database of all Internet content. Spiders and other agents are continually indexing and re-indexing WWW content, but a single World Wide Web site may be visited by an agent once, then not be visited again for months as the queue of sites the search engine must index grows. A site owner can speed up the process by manually requesting that resources on a site be re-indexed, but this process can get unwieldy for large web sites and is in fact, a guarantee of nothing.

Many current internet search engines support two methods of controlling the resource files that are added to their database. These are the robots.txt file, which is a site-wide, search engine specific control mechanism, and the ROBOTS META HTML tag which is resource file specific, but not search engine specific. Most internet search engines respect both methods, and will not index a file if robots.txt, ROBOTS META tag, or both informs the internet search engine to not index a resource. The use of robots.txt, the ROBOTS META tag and other methods of index control is advocated for the purposes of the present invention.

Commonly, when an internet search engine agent visits a web site for indexing, it first checks the existence of robots.txt at the top level of the site. If the search agent finds robots.txt, if analyses the contents of the file for records such as:

```
User-agent: *
Disallow: /cgi-bin/SRC
Disallow: /stats
```

The above example would instruct all agents not to index any file in directories names /cgi-bin/SRC or /stats. Each search engine agent has its own agent name. For example, AltaVista (currently the largest Internet search engine) has an agent called Scooter. To allow only AltaVista access to directory lavstuff, the following robots.txt file would be used:

```
User-agent: Scooter
Disallow:
User-agent: *
Disallow: /avstuff
```

The ROBOTS META tag is found in the file itself. When the internet search engine agent indexes the file, it will look for a HTML tag like one of the following:

```
<META NAME="ROBOTS" CONTENT="NOINDEX,
  NO FOLLOW">
<META NAME="ROBOTS" CONTENT="NOINDEX,
  FOLLOW">
<META NAME="ROBOTS" CONTENT="INDEX, NO
  FOLLOW">
```